图书情报工作
第64卷 第8期 2020年4月

用户交互特征对知识付费行为预测的贡献度研究*

■ 邓胜利 蒋雨婷
武汉大学信息资源研究中心 武汉 429000

摘 要：[目的/意义]为了丰富社交互动行为和知识付费行为的理论研究，有效识别潜在知识付费用户，从而提高在线知识社区的变现能力，在总结前人研究结论的基础上，研究不同类型以及不同程度的交互行为对于预测知识付费用户的贡献度及其变化趋势。[方法/过程]以从知乎社区(www.zhihu.com)爬取的400万用户社会交互行为数据为依托，根据交互主体和交互方向的不同对该社区的用户社会交互行为进行分类，然后，利用随机森林算法研究不同类型和不同程度的交互行为对知识付费用户预测的贡献程度，并对结果进行分析比较。[结果/结论]结果发现，用户和其他用户的交互影响大于用户和平台交互的影响，特别是，用户对其他用户的主动交互行为的影响大于用户接收到的来自其他用户的交互行为的影响。此外，在一定的阈值内，社交互动的程度越大，其对知识付费行为预测的贡献越大。不同的交互类型具有不同的阈值，但是超过这个阈值以后，关系则不再是简单的单调增加关系，可能趋于平缓甚至显著下降。

关键词：社会交互行为 知识付费 用户识别 随机森林 贡献度
分类号：G250
DOI：10.13266/j.issn.0252-3116.2020.08.011

1 引言

随着用户对高质量内容的需求不断增加，知识付费产品由于能降低用户的信息筛选成本，提供实时的优质内容而获得广泛关注。知识付费是指公众利用互联网平台与他人分享自己的认知盈余(指人的智慧、知识、能力、经验等)，从而获得收入的经济现象[1]。当前，大量知识付费产品应运而生，如知乎的"Live"、逻辑思维的"得到""分答"等，这些实时的语音问答互动产品为用户提供了获得特定主题知识以及进行即时互动的平台。以知乎为例，2016年5月推出的"知乎Live"直播服务，鼓励人们通过音频或视频分享他们对特定话题的见解，时长从1小时到2小时不等[2]。用户需要支付一定的费用才能获得进入直播的入口，获得的服务包括获取直播的知识和直播者的互动交流等。

尽管知识付费模式拥有巨大的潜力，但知识付费产品的使用率却远远低于预期[3]。以往的在线知识分享都是基于"共享"模式，这种免费价值观影响了用户为知识付费的意愿。与此同时，知识付费模式也存在一些问题，如知识质量不高、缺乏版权保护等[4]，用户对于知识付费产品的使用率还停留在较低的水平[5]。因此，探究潜在知识付费用户的行为表现特征对于识别潜在的付费用户具有重要的作用。

现有的知识付费行为研究主要集中在知识付费行为的影响因素方面[3-6]，有关潜在付费用户的预测研究仍比较缺乏，仅有的预测研究也只是把直播相关的因素，如直播价格、知识共享者声誉等作为知识付费行为的预测因素，并没有着眼于具体的交互行为类型以及交互程度对于预测潜在知识付费用户的贡献程度，而用户之间的交互行为对于在线社交网站是非常重要的[7]。

笔者旨在探究用户在在线知识社区中的不同类型以及不同程度的交互行为对知识付费行为预测的贡献度，从而为识别潜在付费用户提供有力的依据。笔者试图解决以下两个问题：①哪些类型的用户交互行为

* 本文系国家社会科学基金一般项目"基于用户交互的网络社区知识聚合组织和服务研究"(项目编号：16BTQ072)和国家自然科学基金项目"信息生态链视角下在线知识社区用户贡献行为评价及预测研究"(项目编号：71974149)研究成果之一。
作者简介：邓胜利(ORCID：0000-0001-7489-4439)，教授，博士，博士生导师，E-mail：victorydc@sina.com；蒋雨婷(ORCID：0000-0002-5090-4265)，硕士研究生。
收稿日期：2019-09-17 修回日期：2020-01-29 本文起止页码：93-102 本文责任编辑：徐健

对预测知识付费用户是最有效的;②用户交互程度对预测知识付费用户贡献度的影响。研究结果有助于丰富用户交互行为的理论研究,为用户交互行为与用户知识付费行为之间的关系提供有力的依据。

2 研究综述

2.1 社交网络用户交互行为

人们利用社交媒体来分享他们对不同主题的观点、感受和想法,并在 Facebook 和 Twitter 等社交媒体平台上宣传他们的活动。因此,社交媒体平台提供了大量与人类行为相关的数据,包括社交互动[8]。在线社交互动,又称网络互动,是指用户通过信息交换在网络平台上进行的在线互动行为[9]。网络社交已经成为个人生活的重要组成部分,在互联网时代,它在支持媒体、人与社会之间的交流方面发挥着关键作用[10]。社交网站提供各种各样的社交互动功能,如发布文章、评论、转发、收藏、点赞、关注、分享等[11-12]。以知乎为例,人们与朋友、亲属、同事甚至陌生人建立联系,产生多种社交互动行为。用户可以操纵"关注"他人的状态,"收藏"喜欢的内容、专栏,对他人的内容进行评论、点赞或分享。朋友之间的互动、在线活动的参与、产品推荐和评论都是在网站社区成员之间自由、自愿地编写、发送和阅读。用户所选择的交互类型、在线活动的参与以及产品推荐的购买都是基于自己的选择和偏好。

这些社交互动的交互主体和交互方向并不总是一致的。基于交互主体对社会交互的定义为:"参与主体之间通过一组信息通道(接口)进行的信息交换"[13],为了探讨不同类型的交互行为对付费行为的影响作用,笔者根据主体的不同对交互的类型进行了细分,包括用户与其他用户之间的交互以及用户与平台之间的交互,并绘制了交互类型的关系图(见图 1)。当交互主体是用户与其他用户时,用户 A 会发送信息或者某个状态给用户 B,同时用户 A 也会收到来自用户 B 的信息或状态。当交互主体是用户与平台时,用户 A 会发送信息或状态给平台,而同时也会收到来自平台的反馈。

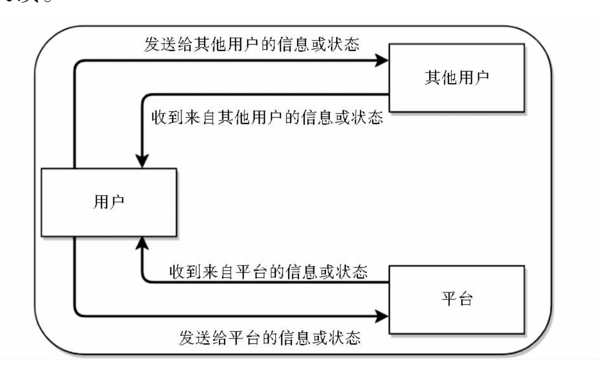


图 1 在线社区用户的社交互动类型的关系

2.2 用户知识付费行为的影响因素

目前大多数研究主要关注用户知识付费行为的影响因素,包括用户自我驱动的因素和在线社区驱动的影响因素。用户自身的影响因素包括信息需求的重要性、信息需求的紧迫性、对付费知识的信任、对知识提供者的认同、对平台的熟悉和信任。社区驱动的影响因素包括:知识质量、价格、社区的访客数量、社区使用便利性。对知识付费行为的影响因素如表 1 所示:

表 1 对知识付费行为的影响因素研究

作者	数据	影响因素	研究发现
(X. Liu & J. Feng, 2018)[14]	使用 Python 爬虫程序从"知乎直播"平台抓取 884 个用户的数据	价格、知识共享者的在线社会资本(粉丝数)、离线社会资本(用户等级)	社交资本对参与者具有显著的影响,对知识消费者支付行为具有正向影响,价格对支付行为没有影响。
(S. Cai, Q. F. Luo, X. Fu, et al, 2018)[15]	使用 Python 爬虫程序从"知乎直播"平台抓取 222 个直播的数据	喜欢直播的用户累计数量、直播的平均评论分数、价格、知识共享者的声誉、性别、对平台的内容贡献程度、以积极回复听众的消息数量	喜欢直播和购买直播的人的数量,交互都会对当前的销售额产生积极的影响。价格对知识共享产品的销售没有影响。
(S. Cai, Q. Luo, X. Fu , et al, 2018)[16]	使用 Python 爬虫程序从"知乎直播"平台抓取 634 个直播的数据	直播的价格、知识提供者的声誉、该知识产品的被点赞数	知识提供者的声誉和该知识产品的被点赞数对销售额有积极的影响。
(张帅, 王文韬, 李晶, 2017)[17]	通过访谈和问卷调查获得了 18 名用户的直播付费意愿及相关信息	个体需求、个体认知、信息质量、主观规范、便利条件、替代品、经济因素	个体需求是用户在线知识付费行为最重要的影响因素,经济因素对知识付费行为的影响相对不重要。
(李钢, 卢艳强, 滕树元,2018)[18]	通过访谈和问卷调查获得了 44 名用户的直播付费意愿及相关信息	感知质量、体验、信任、感知费用、用户资源	付费态度、主观规范、直觉行为正向影响用户付费行为,用户感知内容和 服务质量正向影响用户的付费态度,感知费用与付费态度和付费行为负相关。

大多数研究对付费行为影响因素的研究基于直播相关的特征,如价格、内容质量以及知识共享者的特征,而忽略了用户在平台上的交互行为对其付费意愿的影响。此外,这些研究所依赖的数据样本量是有限的。基于此,笔者基于一个更大量级的、更有说服力的数据集来研究哪些社会交互行为是潜在知识付费用户更有可能在知乎直播平台购买直播的有力信号。

2.3 基于用户交互数据的知识付费行为预测

关于预测知识付费用户的研究表明,用户在社交媒体上的交互行为和交互数据是识别潜在知识付费用户的重要依据。

一些研究对两者关系的机制进行了探讨。研究发现,网络社会互动有利于人际关系的发展,如通过建立信任、建立友谊以及促进人际互动[19]。用户与网络社区的亲密感越强,意味着社区对购买意愿的影响越大。一项基于网络服务的持续意愿研究表明,亲密度和熟悉度影响用户的持续意愿[20]。在社交网络背景下,认知和情感参与,即社交互动,已被证明能够增加用户对朋友推荐的交易的购买意愿[21]。购买意愿是由用户和社交网络社区之间的社交互动直接驱动的。用户在社交网络社区中与他人的社交互动越多,与他人的亲密感和熟悉感就越强。因此,这个用户的购买意愿更有可能受到社区的影响[6]。

还有一些学者为基于用户交互数据的知识付费行为预测提供了有力的依据。H. L. Wu 和 J. W. Wang[22]发现,社交网站所支持的社交互动允许用户培养、增强和维持在线关系,是行为意图的重要预测因素。Y. Zhang 等[23]通过对 Facebook 和 eBay 用户的联合大数据集的研究表明,仅使用社交媒体信息就可以成功预测购买行为,其中,社交网络中用户表现出的兴趣(如 Facebook 的"赞")是预测用户购买行为的重要因素。这些研究表明,利用用户的社交互动行为去识别潜在的付费用户具有科学性和可行性。但是不同类型和不同程度的社会交互对于预测付费行为的贡献度却没有更深一步的研究,因此,笔者拟通过研究哪些交互特征以及哪种交互程度是甄别潜在知识付费用户的最有力信号,以解决上述问题。

3 研究方法

3.1 数据收集

以前的研究是基于问卷调查收集的小样本自我报告数据来衡量的。笔者利用从网站收集的大量累计互动数据作为量化用户社交互动行为的基础,通过 python 编程爬取知乎社区用户的所有行为数据。利用知乎用户的社交关系属性递归爬取了知乎几乎所有存在于关系网中的用户信息。即从一个意见领袖开始,爬取他的关注列表和粉丝列表。再从关注列表和粉丝列表递归地爬取用户的所有行为信息。而散落于关系网之外的用户,由于他们与知乎社交平台中的任一关系网络均无交集,参与度过低,信息量过小,因此不纳入研究范围之内。

截至 2017 年 7 月 10 日,共爬取了 4 376 500 个用户在知乎问答平台上的行为数据。由于数据爬取的时间成本较高,因此笔者假设在数据爬取过程中的数据量变化较小。经过编码、排序、替换缺失值、删除重复值、预测值的类别转换等数据清洗工作,得到了 4 290 000条干净数据。剔除掉用户 id、用户头像链接等无用字段。此外,由于所有用户在商业问题上的变量值都为 0,没有统计学意义,因此笔者删除了商业问题变量,最后,获得实际可用行为字段 17 个,包括 16 类在线社交互动数据以及用户参与的付费直播数。

3.2 用户交互特征及分类

此外,根据交互主体的不同和交互方向的一致性(见图 1),笔者对交互的类别进行了描述。交互行为的分类取决于交互行为的发起者和直接接收者。当行为的发起者是单一用户(研究的本体用户)时,则该交互类型属于用户→其他用户或者用户→平台。若该行为的直接接收者是另一个单一用户(除该用户之外的其他用户),则该交互类型属于用户→其他用户;若该行为的直接接收者是平台上所有的用户或者平台,则该交互类型属于用户→平台。同理,若该行为的发起者是平台,接收者是单一用户,则该交互类型属于用户→平台;若该行为的发起者是除该用户之外的其他用户,接收者是研究的本体用户时,该交互类型属于其他用户→用户,这与平台的功能机制密不可分。变量名称及含义见表 2。

在 17 个特征变量中,有 9 个变量都属于用户对其他用户的主动交互行为,5 个变量属于用户对平台的主动交互行为, 2 个变量是来自其他用户的交互行为,1 个变量属于来自平台的交互行为。在此基础上,笔者对目标变量参加的付费直播数进行了分类处理,见表 3。

3.3 特征重要性(贡献度)的计算

一般情况下,数据集都是高维度的,具有较多的特征,而过多的特征会导致算法的性能以及算法的准确率降低, 因此有必要从某一特定的标准下选择最重要

表 2 知乎的用户社交互动类型、含义及所属类型

特征变量	字段解释	交互类型
voteup_count－赞同数	由其他用户认可的用户答案、文章或专栏的数量	用户→其他用户
favorited_count－被收藏数	用户的内容被其他用户收藏的数量	用户→其他用户
answer_count－回答数	用户在平台上的回答数	用户→其他用户
favorite_count－收藏数	用户在平台上收藏的回答数、话题数、文章数	用户→其他用户
following_count－关注的用户数	用户追随的人数	用户→其他用户
following_topic_count－关注的话题数	用户关注的各个领域的话题数	用户→其他用户
following_columns_count－关注的专栏数	用户关注的个人专栏、话题专栏或独立品牌专栏的数量	用户→其他用户
following_favlists_count－关注的收藏夹数	收藏夹是用户的兴趣收集和知识积累,用户通过关注感兴趣的收藏夹从而获得相关的知识	用户→其他用户
following_question_count－关注的问题数	用户关注的问题的数量	用户→其他用户
question_count－提问数	用户在平台上的提问数	用户→平台
articles_count－文章数	用户在平台上发布的文章数	用户→平台
pins_count－分享数	用户把文章、专栏、回答进行分享的次数	用户→平台
hosted_live_count－主持的付费直播数	“知乎 Live”是一种付费直播产品,知识共享者主要通过语音直播分享自己的知识或者观点给付费的观众。举办的直播数则是用户作为知识共享者主持直播的次数	用户→平台
marked_answers_count－被收录的回答数	用户在平台上的优质回答被知乎收录的次数	平台→用户
follower_count－粉丝数	用户吸引的追随者的数量	其他用户→用户
thanked_count－被感激数	用户在平台上的回答被其他用户感谢的次数	其他用户→用户
目标变量	字段解释	
participated_live_count－参加的付费直播数	用户付费参与的直播数	

表 3 目标变量的分类描述

参加的直播数	类别
=0	非付费用户(类别=0)
>0	付费用户(类别=1)

的特征子集。笔者选择随机森林计算特征重要性。随机森林是一种集成学习算法,原理是将多个弱分类器进行重组,得到高性能的强分类器,该方法能够获得特征贡献值[24]。

3.3.1 随机森林计算特征重要性的原理

随机森林用抽样放回的方法(bootstrap)从样本集中选择 n 个样本组成训练集,用新生成的样本集生成决策树。对于生成的每一个节点,随机地选择 m 个不重复的特征,利用这 m 个特征对新的样本集进行划分,通过基尼系数或者增益率找到最佳的划分依据和特征。假设随机森林拥有 k 个决策树,那么这个过程会重复 k 次。最后,用这个随机森林对预测集进行预测,通过投票获得预测结果。本研究通过基尼指数计算各个特征(社交互动行为的累积量)对目标变量(是否参与付费直播)的贡献。

3.3.2 随机森林计算特征重要性的步骤

笔者通过基尼指数来进行评价,将变量的重要性评分用 VIM(variable importance scores)来表示,基尼值用 GI 表示,假设有 m 个特征 $X_1, X_2, \cdots, X_m$,基尼指数的计算公式如下所示[25]:

$$GI_m = 1 - \sum_{k=1}^{|K|} p_{mk}^2 \quad \text{公式(1)}$$

其中,k 表示有 k 个类别,P_{mk} 表示节点 m 中类别 k 所占的比例。特征 X_j 在节点 m 的重要性由分支前后的基尼指数变化量来表示:

$$VIM_{jm}^{gini} = GI_m - GI_l - GI_r \quad \text{公式(2)}$$

其中,GI_l 和 GI_r 分别表示分枝后两个新节点的基尼指数。

如果特征 X_j 在决策树 i 中出现的节点在集合 M 中,那么 X_j 在第 i 棵树的重要性为:

$$VIM_{ij}^{gini} = \sum_{m \in M} VIM_{jm}^{gini} \quad \text{公式(3)}$$

假设随机森林共有 k 棵树,那么:

$$VIM_j^{gini} = \sum_{i=1}^{n} VIM_{ij}^{gini} \quad \text{公式(4)}$$

最后对所有 VIM 进行归一化处理就得到特征最终的重要性的评分:

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{c} VIM_i} \quad \text{公式(5)}$$

4 交互特征对知识付费行为预测的贡献度分析

在获取的数据集里,付费用户的占比为 9.4%(404 900/4 290 000)。笔者对这批数据进行了统计学的描述,如表 4 所示:

表4　数据集的描述性分析

特征名称	最小值	最大值	均值		标准差
	统计值	统计值	统计值	标准误差	统计值
赞同数	0	3 461 667	565.51	16.823	10 704.853
被收藏数	0	1 087 985	348.53	10.743	6 835.881
粉丝数	0	1 348 183	255.11	9.646	6 137.68
被收录的回答数	0	130	0.03	0.001	0.803
被感激数	0	545 005	118.67	3.235	2 058.594
回答数	0	17 721	14.4	0.121	76.698
提问数	0	3 151	1.71	0.02	12.982
文章数	0	1 534	0.45	0.012	7.64
主持的直播数	0	26	0.01	0	0.177
收藏数	0	239	6.42	0.013	8.129
关注的用户数	0	15 430	98.94	0.317	201.593
关注的话题数	0	22 027	40.27	0.14	89.155
关注的专栏数	0	2 132	11.28	0.044	27.971
关注的收藏夹数	0	5 030	7.63	0.046	29.087
关注的问题数	0	59 975	178.95	0.982	624.854
分享数	0	396	0.22	0.004	2.401
付费参与的直播数	1	1 123	4.04	0.016	10.263

4.1　用户交互类型对知识付费行为预测的贡献度

随机森林有多棵决策树，在决策树中，每个节点都是一个判断条件，根据节点的条件判断结果，这样数据集合将会被分到两个或多个集合中。通过不断划分节点，使得一个分支节点包含的数据尽可能属于一个类别(纯度越来越高)。本文使用信息增益或者基尼杂质来定义纯度。因此，当训练一棵树时，可以计算出每个特征对减少加权杂质的贡献有多大(每个分支节点的数据)。当训练完所有的决策树以后，再对若干棵决策树上杂质的减少进行平均，从而获得最终的贡献值。整个实验过程都通过 python 编程实现，包括数据清洗、数据分块读取和整合、随机森林训练、随机森林预测、以及贡献值和变化趋势可视化。在随机森林的训练过程中，使用了 python 自带的 RandomForestRegressor 接口以简化编程过程。

4.1.1　整体预测值下用户交互类型对知识付费行为预测的贡献度

首先，计算所有用户中用户交互类型对知识付费行为预测的贡献度。笔者训练一个包含 10 000 棵决策树的随机森林来评估 17 个维度特征的重要性。通过观察每个预测器的特征如何影响模型性能来直接度量特征的重要性，对特征的重要性进行了排序，并且用横向条形图进行可视化，结果如图 2 所示：

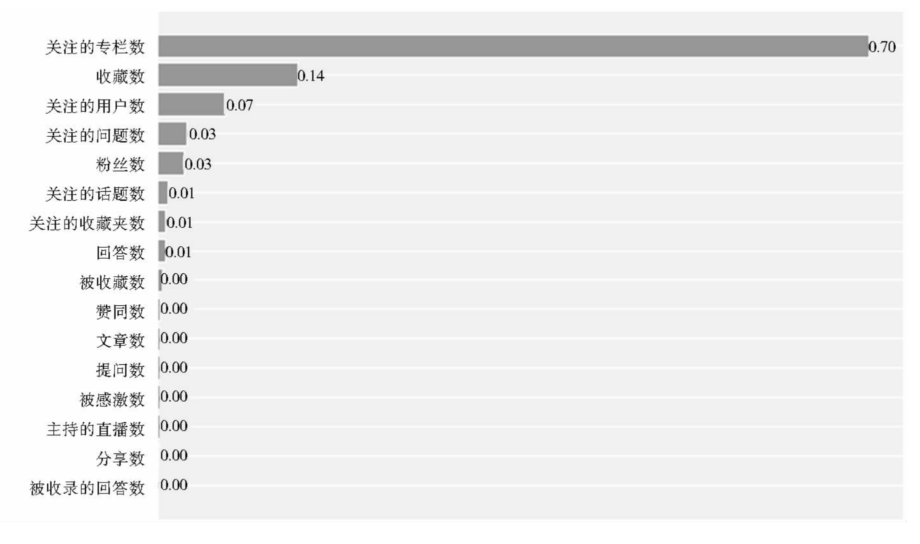


图2　默认特征贡献度

从图 2 看出，对于预测用户是否购买知识付费产品，最重要的 3 个特征为关注的专栏数、收藏数、关注的用户数。基于交互的对象和信息传递方向，这 3 个特征都是属于用户→其他用户交互类型的主动交互行为。

4.1.2 特定预测值下用户交互类型对知识付费行为预测的贡献度

然而，以上的贡献度计算并没有体现特征如何影响特定预测值。因此，笔者根据特定预测值绘制特征的贡献值。本文计算了特定用户组（是否为知识付费）与各个变量的贡献值比较。图 3 表示预测值为 1（付费用户）时各社交互动行为的贡献值，图 4 表示预测值为 0（非付费用户）时各社交互动行为的贡献值。

对于预测付费用户，用户关注的专栏数、收藏数和粉丝数是最重要的 3 个特征，且都是正向的贡献。而关注的用户数和被收藏的次数产生了负向的贡献。

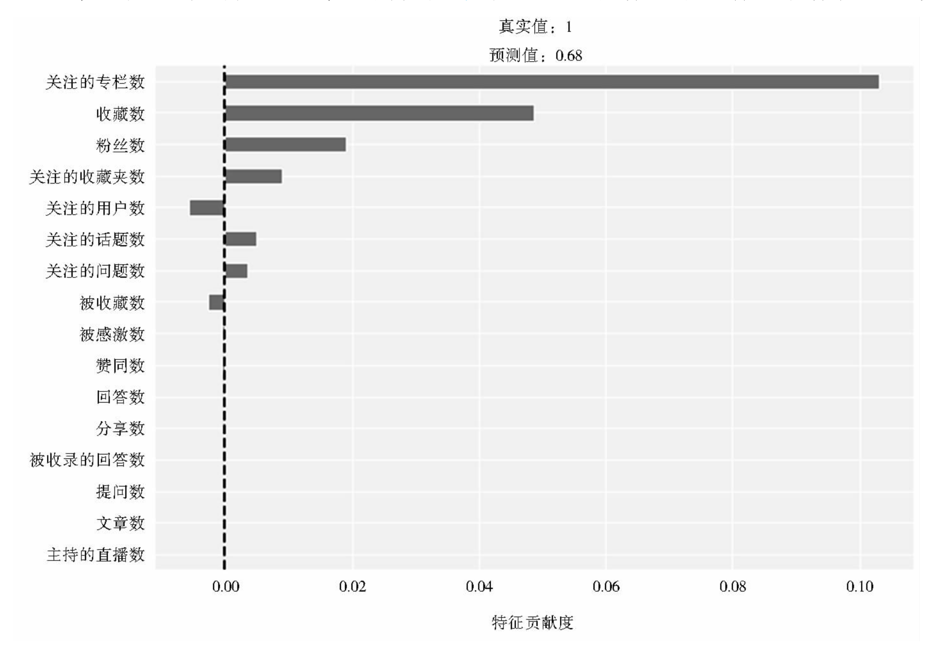


图 3 预测付费用户的特征重要性

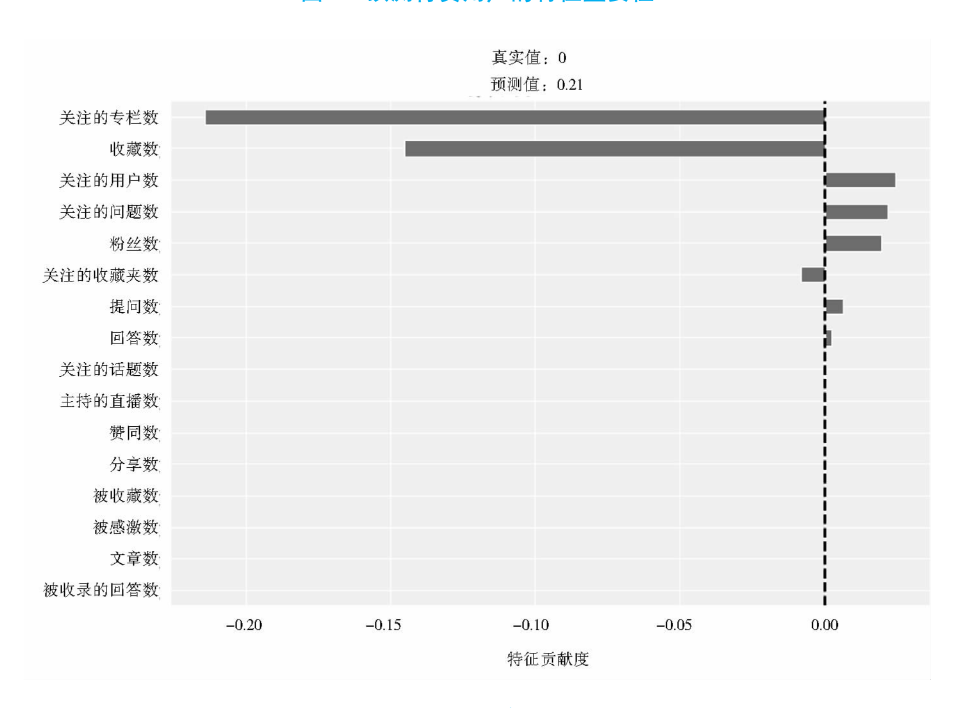


图 4 预测非付费用户的特征重要性

对于预测非付费用户，用户关注的专栏数和收藏数依然是最重要的两个特征，但是用户关注的人数成为第三重要的特征，这和预测付费用户的重要特征不同。此外，最重要的两个特征（关注的用户数和被收藏的次数）对预测有负的贡献，而关注的用户数是正向的贡献。

4.2 用户交互程度对知识付费行为预测的贡献度

以上描述并没有对一个特定的交互行为如何影响用户的知识付费行为有一个全面的解释。因此，笔者分别绘制了预测付费用户（关注的专栏数、收藏数、粉丝数）和非付费用户（关注的专栏数、收藏数、关注的用户数）最重要的前3个特征及其贡献值，以探索社交互动程度和知识付费行为的关系。

由于随机森林本质上是随机的，对于给定的关注专栏数下的贡献具有可变性。然而，平滑的黑色趋势线仍显示出增长的趋势。如图5所示，关注的专栏的数量增加对应于较高的贡献，在贡献达到0.4左右时贡献不再增加。

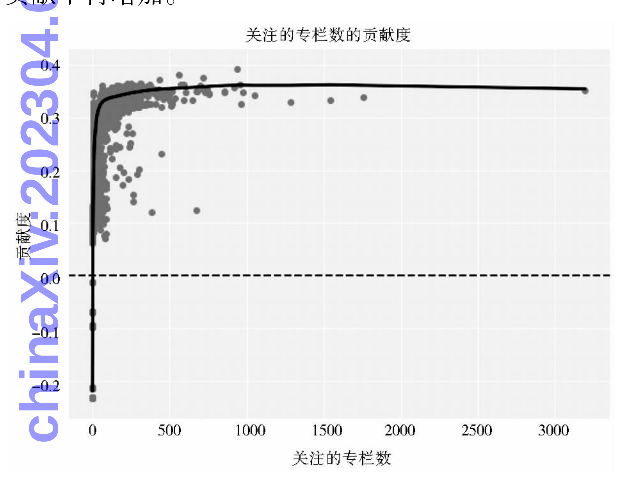


图5 关注的专栏数对付费行为的贡献值

本文绘制了收藏数的贡献值（见图6）。收藏数这个变量的贡献值具有非线性、非单调的特点。低的收藏数具有负的贡献，高的收藏数具有正的贡献。事实上，大部分用户的收藏数量对应一个正的贡献值。当收藏数达到15－50左右时，其贡献值最大，达到0.28。

关注的用户数这一交互行为变量具有复杂、非单调的特点。其贡献值在关注用户数约200左右时达到峰值0.15，此后下降。除此之外，关注的用户数与目标变量似乎具有普遍的负相关关系。

粉丝数这一交互特征对于正向贡献和负向贡献达到了均值，当粉丝数介于1万以内时达到最大贡献，约为0.28的正向贡献和－0.18的负向贡献。

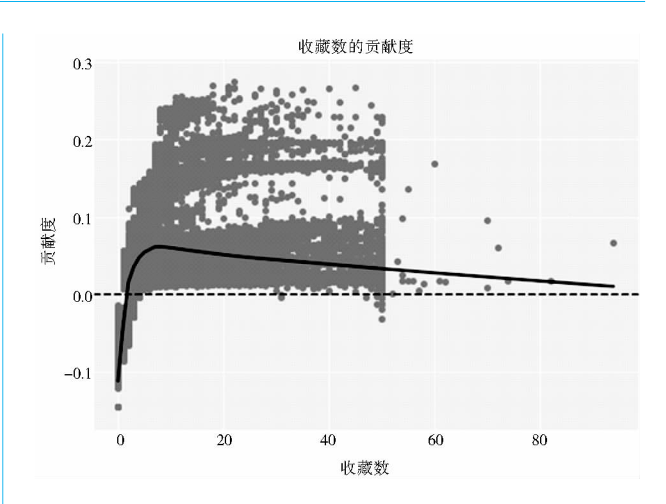


图6 收藏数对知识付费行为的贡献值

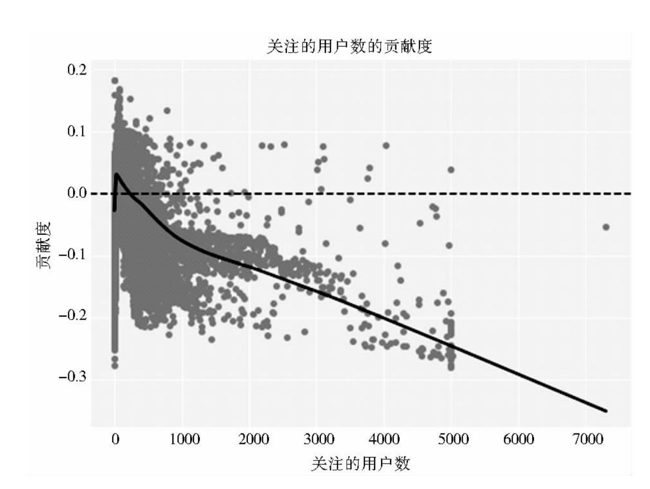


图7 关注的用户数对知识付费行为的贡献值

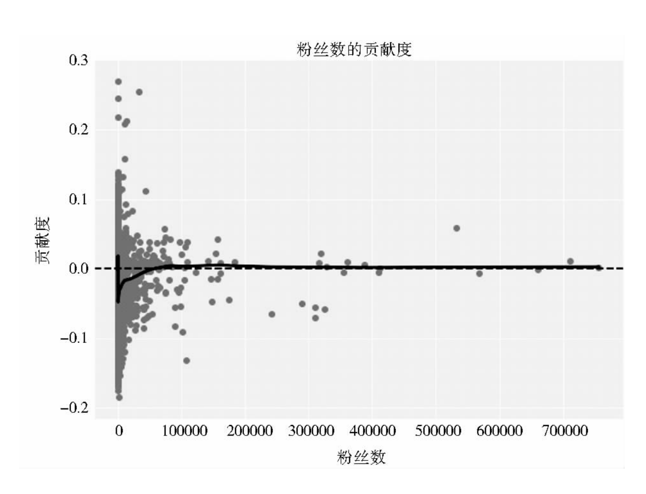


图8 粉丝数对知识付费行为的贡献值

5 讨论

本研究探讨了交互行为的类型和程度对于预测潜在知识付费用户的贡献度大小及变化趋势。

首先，对于交互行为类型的研究发现，关注的专栏数、收藏数、关注的用户数和粉丝数这4种行为的累计

积累值是识别潜在知识付费用户的最有力的信号。在知乎社区中,关注专栏、收藏、关注其他用户这 3 种行为是基于用户兴趣导向的主动行为类型,用户会基于自身的兴趣爱好、自我提升、获取专业知识关注特定主题的专栏,收藏特定主题的帖子以及基于情感因素和社交需求而关注在特定领域内的活跃用户[26-27]。这些行为都是基于用户的个体需求而自发产生的行为,因此,这些用户需求导向型的交互行为都是体现用户个体需求的有力指标,而个体需求是用户为产品付费的最重要的驱动因素之一[28]。已有学者对通过访谈获得的数据进行定性研究发现个体需求是知识付费行为最重要的特征[17]。笔者从定量的角度入手,通过对大数据的挖掘,采用机器学习算法获得了和基于定性方法研究相同的结论。研究结果表明,当用户个体需求越大,包括基于兴趣的关注越频繁,或者兴趣范围越广泛,其可能参与付费直播的可能性就越高。此外,笔者从亲密感的角度解释粉丝数量对于知识付费行为的影响。粉丝数量是用户声誉的重要体现指标,较高的粉丝数量对于增强用户的社区归属感和亲密感有重要作用,而这种亲密感会显著影响用户的付费行为,这与之前的研究是一致的[20],即亲密度和熟悉度影响用户的持续购买意愿。此外,先前的研究表明,用户与他人的亲密感和熟悉感越强,这个用户的购买意愿更有可能受到社区的影响[6]。

其次,笔者在研究社会交互程度对知识付费用户的贡献度方面有了一些发现。之前的研究表明交互程度越深,用户知识付费意愿越强[3]。但笔者研究发现社会交互程度和知识付费行为并不是单一的伴随增减关系。在一定范围内交互程度和特征重要性存在伴随增加的关系,一旦超过某个阈值,这个信号就保持平稳不变甚至减弱。也就是说,只有在一定的阈值内,用户与平台的交互程度越多,对平台越熟练,其为知识付费的可能性更大。笔者为这一研究结果提供了可能的解释。关注的专栏数、关注的用户数和收藏数这些主动交互行为的交互程度和特征重要性在从 0 开始的某个区间里有明显的增幅,达到阈值后,呈显著下降趋势或者平稳不变。主动交互行为的增加在一定范围内能够显著提高用户进行知识付费的可能性,但是对于交互程度较深的用户,这种提升作用非常弱,由于交互程度很深的用户多为意见领袖,而这类用户主要扮演知识分享的角色,而很少作为知识付费用户从直播中获取知识,因此这部分用户并不是最佳的商业变现的潜在用户群体。而对于粉丝数这个交互特征而言,在从 0 开始的某个范围内,特征重要性负向增加,达到最大的负向贡献以后,贡献度趋近于 0。粉丝数作为一个用户的个人声誉,对于用户知识付费行为的促进作用是有限的,并且在一定范围内,粉丝数越多,其为知识付费的可能性就越小。同样,这是由于粉丝数较多的用户多为意见领袖,意见领袖的交互行为多以知识贡献为主。因此,粉丝数的增加对于用户知识付费行为的预测在一定范围内是负向的促进作用。

6 结论与展望

笔者基于随机森林算法研究了社会交互行为类型以及交互程度是潜在知识付费用户预测的重要因素。

首先,就交互类型而言,预测付费用户和非付费用户时,特征具有不同的重要性。在预测付费用户时,关注的专栏数、收藏数、粉丝数对于预测的贡献值最大,而在预测非付费用户时,关注的专栏数、收藏数和关注的粉丝数的影响最大。因此,关注的专栏数、收藏数、关注的用户数和粉丝数这 4 个交互特征对于知识付费行为的贡献是最大的,由表 2 可知,关注专栏、收藏以及关注其他用户这 3 类行为都属于用户→其他用户类型。而粉丝数则属于其他用户→用户类型。这 4 类行为都是用户和其他用户的交互,因此研究表明用户跟其他用户之间的交互行为对于预测的贡献度远大于用户和平台的交互。此外,在这 4 类行为中,只有一类行为属于来自其他用户的交互,也就是说,用户的主动行为相比于用户的被动行为,其对于潜在知识付费用户的预测贡献度更大。而在所有的交互行为中,个体需求驱动的交互行为对于付费用户的识别发挥着最重要的作用。该研究提供了一些新的理论见解。首先,从定量分析角度证实了前人通过定性分析方法的研究结论,即个体需求,包括兴趣爱好、自我提升、情感需求和社交需求是用户购买知识付费产品最重要的驱动因素。其次,根据交互行为的定义细化了交互的对象和交互路径,对于社交网站上用户交互行为的研究提供了新的研究视角。

同时,该结果为在线知识平台识别潜在的知识付费用户提供了有效的实践意义,结果发现在知乎社区中,专栏、收藏和关注的用户的行为应该被作为主要的预测指标,这些主动交互行为作为最重要的识别特征,可有效减少平台进行特征筛选的时间。此外,利用随机森林算法,对不同交互特征类型的重要性进行了定量分析,这种方法在海量用户数据集中的适应性得到了有效验证。

此外,就交互程度而言,笔者发现,交互程度和交互行为的关系并不总是呈现简单的单调增减的关系。交互程度在一定范围内的增加对于付费行为的影响有显著提升,不同类型的交互行为具有不同的阈值,而超过某个特定范围时,这种影响趋于平缓甚至有下降的趋势。研究结果表明,交互程度越深的用户并不总是潜在的知识付费用户。该研究结果从理论层面为一些研究结论提供了一些必要的限制条件。之前的研究认为交互程度越深,其越有可能为知识付费。但是这在知乎社区并不是普遍存在的现象。交互程度在一定范围内的增加是有效的,但是超过一定范围,这种交互程度的增加对于识别潜在的知识付费用户的重要性趋于平稳甚至显著下降。从实践层面考虑,对于平台而言,监测用户的累计交互程度是一种有效的识别潜在知识付费用户的途径。根据研究结果,当某个交互特征的程度处于特定区间时,其为知识付费的可能性是最大的,对交互程度的监测有助于平台进行定向的直播付费产品投放。

与现有相关研究相比,本研究的创新点主要体现在4个方面:①在理论上,从用户交互行为的视角识别潜在的知识付费用户,并且阐述了利用用户在线交互行为识别潜在知识付费用户的理论依据。②使用了大量级的真实用户交互数据作为实验依托,400万的用户交互数据使得实验结果更具科学性和可信性。③考虑了知乎社区上几乎所有的交互行为。笔者剔除掉了没有统计意义的行为字段,其他字段全部保留,使得研究结果更加完善。④使用了科学的研究方法,运用随机森林算法对大量级的数据进行特征的重要性计算。此外,利用python框架进行了有效的可视化表达,以提高文章的可读性。

当然,本文也有一定局限性。笔者仅从一个在线社区入手,研究的特征变量不具有广泛的适应性,未来的研究方向应通过探索具有不同文化背景、不同运营模式[29]的在线知识社区以获得具有可比较的行为特征变量的重要性。在此基础上,所在行业的特征也应该被考虑。

参考文献:

[1] 中国互联网协会分享经济工作委员会国家信息中心分享经济研究中心. 中国分享经济发展报告2017[EB/OL].[2019－08－26]. http://www.sic.gov.cn/archiver/SIC/UpFile/Files/Htmleditor/201703/20170302125144221.pdf.
[2] LI G. Exploring users' motivation to contribute in online platforms[EB/OL].[2019－08－26]. http://laitinej.mbnet.fi/jouni/Artikkelit/DC_paper_4.pdf.
[3] NG C S P. Intention to purchase on social commerce websites across cultures: A cross-regional study[J]. Information & management, 2013, 50(8):609－620.
[4] 周涛, 檀齐. 基于社会资本理论的知识付费用户行为机理研究[J]. 现代情报, 2017(11):48－52.
[5] 李武. 在线知识付费平台:何去何从?[J]. 图书情报知识, 2018, 183(3):2.
[6] ZHAO Y, LIU Z, SONG S. Why should I pay for the knowledge in social Q&A platforms? [C]//International conference on information. The United Kingdom:Springer, Cham, 2018:577－582.
[7] DENG S, TONG J, LIN Y, et al. Motivating scholars' responses in academic social networking sites: an empirical study on ResearchGate Q&A behavior[J]. Information processing & management,2019, 56(6):102082.
[8] SCHOEN H, GAYO-AVELLO D, TAKIS METAXAS P, et al. The power of prediction with social media[J]. Internet research, 2013, 23(5):528－543.
[9] CHEN C C, LIN Y C. What drives live-stream usage intention? the perspectives of flow, entertainment, social interaction, and endorsement[J]. Telematics and informatics, 2018, 35(1):293－303.
[10] BLAZEVIC V, WIERTZ C, COTTE J, et al. GOSIP in cyberspace: conceptualization and scale development for general online social interaction propensity[J]. Journal of interactive marketing, 2014, 28(2):87－100.
[11] ALLEB S M, CHORLEY M J, COLOMBO G B, et al. Exploiting user interest similarity and social links for micro-blog forwarding in mobile opportunistic networks[J]. Pervasive and mobile computing, 2014, 11:106－131.
[12] LU X, YU Z, GUO B, et al. Predicting the content dissemination trends by repost behavior modeling in mobile social networks[J]. Journal of network and computer applications, 2014, 42:197－207.
[13] STORRS G. A conceptualisation of multiparty interaction[J]. Interacting with computers, 1994, 6(2):173－189.
[14] LIU X, FENG J. Research on the influencing factors of the willingness to pay for knowledge consumers in the knowledge payment platform[EB/OL]. [2019－08－27]. https://aisel.aisnet.org/cgi/viewcontent.cgi? article=1016&=&context=whiceb2018&=&sei-redir=1.
[15] CAI S, LUO Q F, FU X, et al. Paying for Knowledge: Why people paying for live broadcasts in online knowledge sharing community? [EB/OL].[2019－08－27]. https://aisel.aisnet.org/pacis2018/286.
[16] CAI S, LUO Q, FU X, et al. Paying for live broadcast: predicting internet knowledge product sharing[EB/OL].[2019－08－27]. https://pdfs.semanticscholar.org/ab6b/210302138823e79b39e30a118a50f1fadec5.pdf.
[17] 张帅, 王文韬, 李晶. 用户在线知识付费行为影响因素研究

[J]. 图书情报工作, 2017,61 (10):94 – 100.
[18] 李钢, 卢艳强, 滕树元. 用户在线知识付费行为研究——基于计划行为理论[J]. 图书馆学研究, 2018, 429(10):39 – 50.
[19] BOCK G W, ZMUD R W, KIM Y G, et al. Behavioral intention formation in knowledge sharing: examining the roles of extrinsic motivators, social-psychological factors, and organizational climate [J]. MIS quarterly, 2005, 29(1):87 – 111.
[20] LEE Y, KWON O. Intimacy, familiarity and continuance intention: an extended expectation-confirmation model in web-based services[J]. Electronic commerce research and applications, 2011, 10(3):342 – 57.
[21] SHIN J K, PARK M, JU Y. The effect of the online social network structure characteristics on network involvement and consumer purchasing intention: focus on Korean social promotion sites[C]//The 11th International DSI and the 16th APDSI Joint Meeting, Taipei: Asia Pacific Decision Sciences Institute, 2011.
[22] WU H L, WANG J W. An empirical study of flow experiences in social network sites[EB/OL]. [2019 – 08 – 27]. https://aisel.aisnet.org/pacis2011/215.
[23] ZHANG Y, Pennacchiotti M. Predicting purchase behaviors from social media[C]//Proceedings of the 22nd international conference on World Wide Web. Brazil: ACM, 2013:1521 – 1532.
[24] 徐少成, 李东喜. 基于随机森林的加权特征选择算法[J]. 统计与决策, 2018(18):25 – 8.
[25] 姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报(工学版), 2014, 44(1):137 – 141.
[26] 周涛, 檀齐, BAYAN T, 等. 社会交互对用户知识付费意愿的作用机理研究[J]. 图书情报工作, 2019(4):94 – 100.
[27] WANG J J. A research on the development trend of knowledge payment based on Zhihu. I[M].//LI Zhiqing. The new silk roadlLeads through the Arab peninsula: mastering global business and innovation. Bingley: Emerald Publishing Limited, 2019: 229 – 241.
[28] 李武, 艾鹏亚, 许耀心. 在线付费问答平台的用户付费模式及付费意愿研究[J]. 图书情报工作, 2018, 62(13):24 – 29.
[29] 吴楠. 从知乎 Live 看知识付费[J]. 传播力研究, 2019, 3(4):217.

作者贡献说明:
邓胜利: 确定论文选题, 提出整体研究思路和框架, 修改定稿;
蒋雨婷: 负责研究设计, 数据收集与处理, 撰写与修改论文。

Research on the Contribution of User Interaction Characteristics to the Prediction of Knowledge Payment Behavior

Deng Shengli Jiang Yuting
School of information management, Wuhan University, Wuhan 429000

Abstract: [Purpose/significance] In order to enrich the theoretical study between social interaction behavior and knowledge payment behavior, identify potential paying customers effectively so as to improve the marketability of online knowledge community, this paper studies the contribution degree of different types and degrees of interactive behaviors to the prediction of knowledge paying users and their changing trends on the basis of summarizing the predecessors' research conclusion. [Method/process] Based on the social interaction behavior data of 4 million users crawled from the Zhihu community (www.zhihu.com), this paper classified the social interaction behavior of users in the community according to the different interaction subjects and interaction directions, and then used random forest algorithm studied the contribution of different types and degrees of interaction behavior to the prediction of knowledge paying users. [Result/conclusion] The results show that the impact of the interaction between users and other users is greater than that of the interaction between users and the platform. In particular, the impact of active interactions sent to other users is greater than the impact of passive interactions from other users. In addition, within a certain threshold, the greater the degree of social interaction, the greater its contribution to the prediction of knowledge payment behavior. Different interaction types have different thresholds, but beyond this threshold, the relationship is no longer a simple monotonic increase and may tend to be flat or even significantly lower.

Keywords: social interaction behavior knowledge payment user identification random forest contribution degree